STATISTICS WORKSHOP II

United States Department of Agriculture

Exploratory and Confirmation Data Analysis

presented by
Vicki Lancaster
NEPTUNE AND CO., INC.
vlancast@neptuneandco.com

What is data,

> *A collection of numerical values recording the magnitudes of various attributes of the objects under study.*
>
> (Hand, 1999)

What is data analysis,

> *The processing of the data.*
>
> (Hand, 1999)

Why do we need data?

In God we trust,

all others must bring **data**.

Unknown

Data Analysis *Introduction*

DataData analysis is *not* a case of simply applying a directorydirectory of tools to adirectory of tools to a given prob

critical assessment,
exploration,
testing, and
evaluation.

ItIt is a domainIt is a domain the requires *intelligence* and and *c* asas the application of *knowledge* and and *expertise* about the data. It is a challenging and demanding discipline.

It is a discipline that which is continuing to evolve.

(Hand, 1999)

Data Analysis *Introduction*

ThereThere are two broThere are two broadThere ar *exploratory* and *confirmatory*.

1. *ExploratoryExploratory data analysis* (EDA) is concerned w **searching** for clues and finding evidence.

2.2. *Confirmatory data aConfirmatory data analysis* (CDA) (CDA) i **evaluating** the evidence.

SESSION OUTLINE

Data Analysis

EDA	**CDA**
Four Themes of EDA	*Goodness-of-Fit Tests*
1. resistance	*1. chi-square*
2. residuals	*2. EDF*
3. re-expression	*3. moment*
4. displays	*4. regression*

EDA *Introduction*

What is *exploratory data analysis* (EDA)?

EDAEDA is a process that uses non-EDA is a process that u suchsuch assuch as graphical methods, to gain insight into asu data.

I t It character It characterizes It characterizes It characterize and modeling are driven by data.

IfIf you If you think If you think of your data set as a sto numbers,numbers, tnumbers, thennumbers, then EDA is the storystory writstory writtenstory written in numbers to pictures.

EDA methods are used:

to *isolate* patterns and relationships,

to *uncover* unexpected behavior,

to *confirm* or *disprove* or assumptions, and

to *reveal* information.

Why is *exploratory data analysis* important?

MostMost classical procedures are based on *asassuassumpti aboutabout the characteristics* of a of a variable, and of a variable, thethe analyses depends upon *the validity o the vali assumptions*.

TheThe *graphical methods* of EDA provide powerful diagnosticdiagnostic toolsdiagnostic tools for confirming assu assumptionsassumptions are not met, for suggesting corre actions.

For example,

ifif you if you bif you blindly conducted a *one sample t*-test that looked like ...

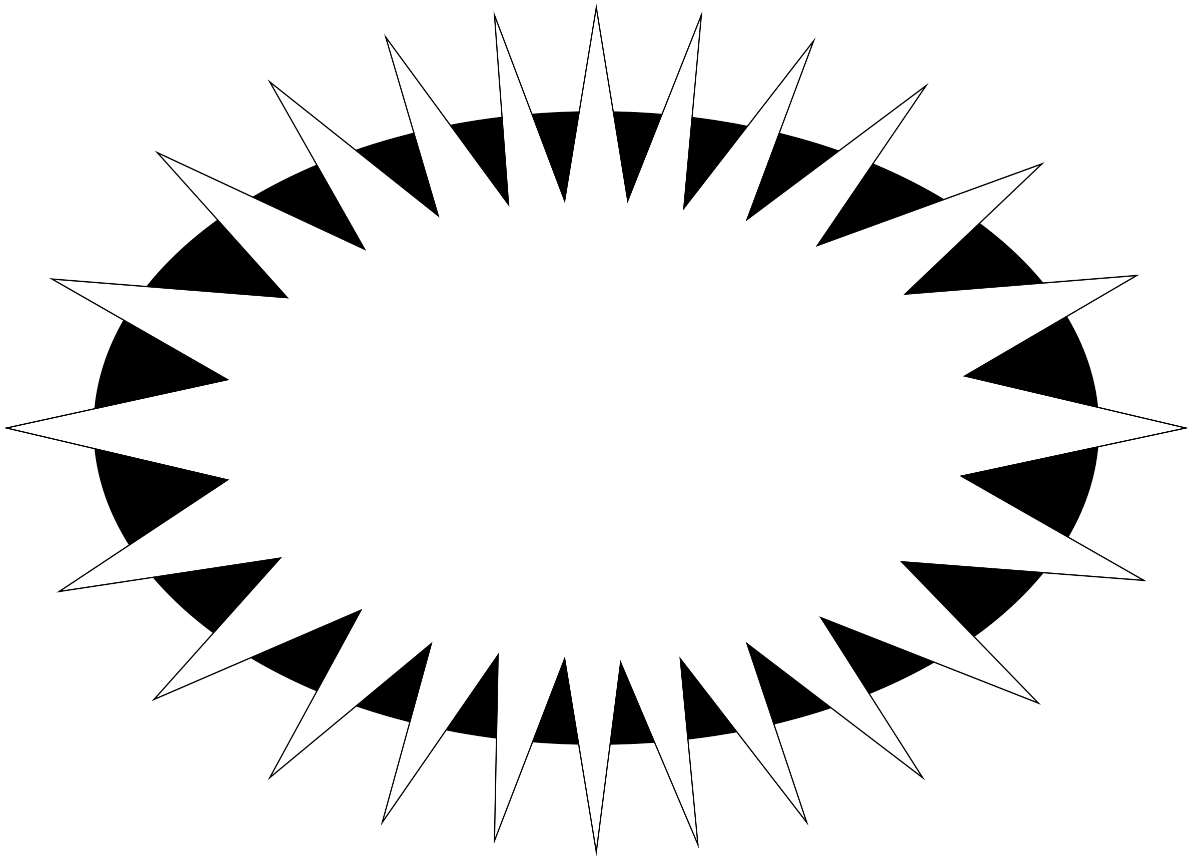
you would fail to reject the null hypothesis.

But,
if you had done a few EDA plots on the data first ...

you would have noticed a potential outlier.

If,
itit turnsit turns out the outlierit turns out the outlier is due to
example a data entry error) and it is corrected ...

The results of your t-test are now different.



Graphical Methods for Data Analysis
John M. Chambers
William S. Cleveland
Beat Kleiner
Paul A. Tukey

TheThe firstThe first published presentation of EDA (1970 - 19 waswas the preliminwas the preliminawas the preliminary Tukey. His 1977 book,

Tukey,Tukey, J. W. (1977).Tukey, J. W. (1977). *ExpEx* Addison-Wesley, Reading Mass.

representsrepresents the definitive account on the subjerepre wwantewantedwanted to dispel the *myth* that we are not allo look at the data prior to modeling.

AtAt thaAt thatAt that time there was a tension betw competing points of view:

that a hypothesis must not be data driven; and

thatthat EDA wasthat EDA was needed prior to inferential toto understand hoto understand how rich to understar support.

Four Themes of **EDA**

Resistance

Residuals

Re-expression

Displays

Four Themes of EDA

Resistance is is is a term used to denote a property of measmeasuresmeasures measures of location or spread relativelyrelatively unaffected byrelatively unaffected by the medianmedian is amedian is an example of a resistant locationlocation and the interquartile range (IQR) is an example of a resistant measure of spread.

A summary statistic is *resistant* if

itit is insensitive to anyit is insensitive to any small change of the data, and

to any large change in a small part of the data.

Four Themes of EDA

Robust versus Resistance

Robust is used to describe an *inference procedure* that that is stablestable when model assumptions are violatstable whe exexaexample,example, the *t-test* is *robust* with respect assumption of normality.

Robustness sensitivity to model assumptions.

Resistance is used to describe a *sstatiststatistic* that is arithmeticallyarithmetically stablearithmetically stable under datadata values. For example, the *median* is a *rresisresista* estimator.

Resistance sensitivity to the data.

Resistant summary statistics:

paypay attention to the *main bodmain body* of the data given little attention the outliers; and

areare useful in graphical methare useful in graphical r the construction of box plots.

For example,For example, look at theFor example, look at the 9 and 10. Are any of these *resistant*?

Four Themes of **EDA**

EDAEDA charactEDA characteriEDA characterized decomposing the data into structure and noise,

$$\textit{data} = \text{fit} + \textit{residuals},$$

andand and thenand then and then examiningand then examin movemove it into the fit. The fitting process would then be repeatedrepeated and frepeated and forepeated and f analysis.

ThisThis process has its roots in tThis process has its roo paradigmparadigm of partitioning variabilityparadigm of pa parts,parts, explained and parts, explained and unparts, explai notionnotion simply usnotion simply uses tnotion simply us thatthat only on the obthat only on the observedthat on treatment possible.

Four Themes of **EDA**

TheThe philosophy of EDA is thaThe philosophy of EDA is t datadata isdata is not complete without a careful examination the *residuals*.

ResiResistantResistant Resistant analyResistant analyses provi dominantdominant behavior and unusual behavior in dat

Residuals contain any contain any drastic departures cont pattern, as well as random fluctuations.

Four Themes of **EDA**

Re-expression i involves the question of what scale would help to simplify the analysis of the data.

Re-expression into another scale may help to

achieve symmetry,
facilitate interpretation,
promote constancy of variance,
achieve a more linear relationship, or
simplify structure for two-way tables.

depending on the structure of the data.

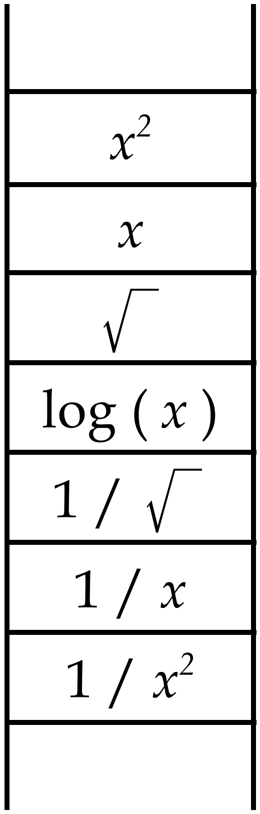
Re-Re-expRe-expression most often comes from the family o functionsfunctions known functions known afunctions kno take y into y^n, together with the logarithm.

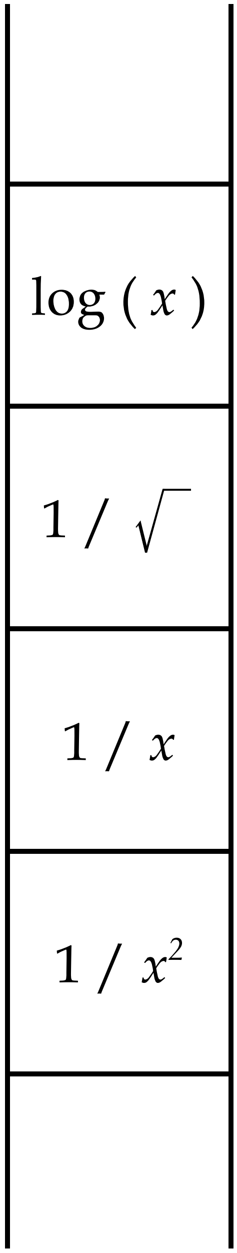
Ladder of Transformations: $y = x^p$

$p =$	*2*	x^2	square
$p =$	*1*	x	(none)
$p =$	*1/2*	$\sqrt{\ }$	square root
$p =$	*0*	log (x)	log
$p =$	*1/2*	1 / $\sqrt{\ }$	reciprocal square root
$p =$	*1*	1 / x	reciprocal
$p =$	*2*	1 / x^2	reciprocal square

Four Themes of EDA

InIn In conjuIn conjunction wiIn conjunction withIn conjunctio *plot* t the ladder of transformat the ladder of suggestionssuggestions for transforming to achieve norma





log (x)

1 / $\sqrt{}$

1 / x

1 / x^2

Four Themes of **EDA**

Displays meet the need to see the meet the need to see the beha
toto reveal the unexpected features, such as *outliers*; and
confirmconfirm or disprove *assumptions*, such as the
distributional assumptions of normality.

TukeyTukey argued that (Tukey argued that (goTukey argu
allowallow unexpected valuesallow unexpected values to pre
identifiedidentified,identified, identified, models can be ex
account for them.

What are *outliers*?

AnAn *outlier* is defined as any value of the is defined as variablevariable that falls outside the pattern of the other values.values. Exacvalues. Exactlyvalues. Exactly wha subsubjecsubjective.subjective. There is no quantitative rul thatthat can be generalized to allthat can be generalize identifying outliers.

Why is the detection of *outliers* important?

OutliersOutliers can greatly influence the value Outliers statisticstatistic andstatistic and the costatistic and the conclu

InIn many instances there isIn many instances there is an assi outlier(s) such as an input error.

ButBut even when there is no assignaBut even when there outlieroutlier and it isoutlier and it is aoutlier and it is a true meanmean the data value is not vamean the data value changed or omitted.

Kruskal (1960a) *doctrine* states:

> i t it is of it is of great importance to preach the it is of great importance to pr outliersoutliers should <u>always</u> be reported, even be reported, even when one theirtheir causes aretheir causes are known or when one rejects them fortheir ca goodgood rule or reason. The immediate pregood rule or reason. The im statisticalstatistical analysis are almstatistical analysis are almosstatistical a suppressingsuppressing announcement of observations thatsuppressing anno pattern;pattern; we must maintain a stpattern; we must maintain a strpat pressures.

Displays of **EDA**

TThereThere aThere are two principal methods for dealin ououtliers:outliers: *identification* and *accommodation*. . outlier(s)outlier(s) is detectedoutlier(s) is detected or identifie one of several ways (Tietjen, 1986):

omitomit the outlier(s) and treatomit the outlier(s) and trea a new sample;

omitomit the outlomit the outlier(somit the outlier(s) censored sample;

askask the experimenter fask the experimenter forask th to replace (verify) the outlier;

WinsorizeWinsorize the outliers, Winsorize the outliers, i value of the nearest good observation;

presentpresent all present all analyses with and outlier(s).outlier(s). outlier(s). When there are no differen thethe two sets of analyses the dilthe two sets of analyse WhenWhen there is a differencWhen there is a differe bobothboth results should be pboth results sh conclusions.

Displays of **EDA**

What are *assumptions*?

Assumptions are the rules under w are the rules conclusionsconclusions drawn from applying an inferen method are valid.

TheThe implicitThe implicit *assumptions* of an inferential are the rules that govern its use.

TheseThese methods are trusThese methods are trustw rules that govern their use are met.

Displays of EDA

Why is it important to evaluate *assumptions*?

TheThe distributional assumptions of the data will inThe distr determinedetermine what inferential statistical method appropriate, parametric or non-parametric.

EstimationEstimation procedures such as the calculatiEst confidconfidenceconfidence intconfidence intervals, tc predictionprediction iprediction intervals depend str underlyingunderlying distribution. When a distribution assumedassumed extreme tail percassumed extreme tail pe thisthis is this is important in environmental work where da are often lognormal.

***Displays of* EDA**

TheseThese *assumptions* can be viewed can be viewed as fal categories:

1.1. those that constrainthose that constrain *howhow the data are* asas the requirement of a random saas the requiremen andand and theand the use of an appropriate rando processprocess foprocess for assigning treatment (determined by the experimental design); and

2. thosethose that constrain thethose that constrain the *cha* suchsuch assuch as a reqsuch as a requirement the distributed.

AnAn experimental design is adopted *prioprior* to data collecticollectioncollection tcollection to assure the result generageneratedgenerated by a random (or a restricted randon process.

EDAEDA techniques are employedEDA techniques are emplo collected to evaluate the characteristics of the data.

Displays of **EDA**

AA majorA major contributionA major contribution of the de withwith EDA has been the emwith EDA has been the em and the variety of new graphical techniques.

CommonlyCommonly used plots to check for distributio assumptions and outliers include,

histograms,

box-plots, and

quantile - quantile plots (*Q-Q plots*).

Displays of **EDA**

Definitions . . .

AA *boxplot* is a rectangle, the is a rectangle, the top and bo rectanglerectangle representrectangle represent the upper ofof the data, the hof the data, the horizof the rectrectanglrectanglerectangle represents the median. shapeshape of a T ,shape of a T , extend fromshape of a valuevalue not beyond a *standard span* fr from t quartiles.quartiles. These lines are oftquartiles. These whiskers.whiskers. whiskers. Values beyond the end ofw are drawn individually.

TheThe *standard span* is 1.5 InteInter-Quartile Range (IQR).

Definitions . . .

TheThe *quantile* of the dat of the data is of the data is a
thethe datthe data intthe data into two groups, so tha
observationsobservations fall belowobservations fall belo
fallfall above the quantile. For example, the 75th
qquantilequantile (Q(.75)) divides the data set suchqua
threethree fourths ofthree fourths of the observations fall b
and one fourth fall above.

Note: The *median* is the 50th quantile, Q(.50).
TheThe *upper upper quaupper quartile* is the 75th quan
TTheThe *lower quartile* is the 25th quantile, Q(.25).
The *IQR* = Q(.75)   Q(.25).

Figure 1. Data from a
Normal Distribution, Sample Size=38

Figure 2. Data from a
Lognormal Distribution, Sample Size=38

Displays of **EDA**

TheThe box plot is a visual displayThe box plot is a visual disp
the *five-number summary* of a data set.

Definition . . .

> TheThe *fivfive-numbefive-number summary* of a data set co
> thethe the smallestthe smallest the smallest observthe smal
> thethe lower quartilethe lower quartile (bottomthe lower q
> (lin(line(line in the box), the upper quartile (top of (lin
> box),box), andbox), and the largest observation (upperbox
> whisker),whisker), written inwhisker), written in order fro
> largest.

ForFor example, the *five-numberfive-number summary* for the d
displayed in Figure 1 is:

	Med.	
Q(.25)		***Q(.75)***
Min.		***Max.***

	4.30	
3.30		6.45
0.51		8.09

ItIt is iIt is inIt is interesting to compare the descriptive statistics the data displayed in Figures 1 and 2.

Descriptive Statistics	***Fig. 1 Normal***	***Fig. 2 Lognormal***
Lower Quartile	3.30	33.72
Mean	4.68	768.50
Median	4.30	168.70
Upper Quartile	6.45	573.20
Standard Deviation	2.06	1828.52
Range [Min. Max.]	7.58	10116.74
IQR [Q(.75) Q(.25)]	3.15	539.48
Coefficient of Skewness	0.14	4.08
Coefficient of Kurtosis	2.12	20.13

Definitions . . .

TheThe third moment about the mean is aThe third mome
aasymmetryasymmetry called *skewness*. Symm.
distributionsdistributions will have a skewness o
distributionsdistributions that aredistributions that are sk
aa skewnessa skewness < 0, and distributions that area sk
to the right will have a skewness > 0.

OfOf interest is the *standardized third moment* or the *coefficient of skewness,* $\sqrt{\quad}$,

where,

$$\sqrt{\quad} \qquad ,$$

$$\bar{\ } \qquad , \text{ and}$$

$$\bar{\ } \qquad .$$

Definitions . . .

TheThe fourth moment about tThe fourth moment about ofof cof curvatof curvature or *kurtosis*, which is the de flatness of a density near its center.

OfOf interest is the *standardized fourthstandardized fourth m coefficient of kurtosis*, b_2,

where,

,

–

, and

–

.

Displays of EDA

ValuesValues of $\sqrt{\ }$ and b_2 close to 0 and $3(n \quad 1)/(n + 1)$, respectively indicate normality.

VaValuesValues differing from these are indicators of noV normality.

TheThe signs The signs anThe signs and magnitude of thes information on the type of non-normality.

$\sqrt{\ }$ > 0 positively (or right) skewed,

$\sqrt{\ }$ < 0 negatively (or left) skewed,

$b_2 > \ 3(n \quad 1)/\ 1)/(n + 1) + 1)$ relates to heavier tails than the normal, and

$b_2 < < \ 3(n \quad 1)/(n + 1)$ relates to lighter tails than the normal.

Displays of **EDA**

SomeSome observations on theSome observations on the desc
Normal distribution:

mean median,

skewness 0,

kurtosis 3.

SomeSome observations on the descriptiveSome observations
Lognormal distribution:

mean >> median,

skewness >> 0,

kurtosis >> 3.

Definitions . . .

AA *histogram* partitions partitions t partitions the range severalseveral nonoverlappingseveral nonoverlapping in calledcalled bins, and counts thcalled bins, and counts th inin each bin. Thein each bin. The number of counts in ea bebe displayed on a densitbe displayed on a density scbe representsrepresents the prorepresents the probabirep frequencyfrequency scale, where thfrequency scale, whe binbin counts. The histogram is completely determineddetermined by two parameters, the *binbin wid* the *bin origin*.

Note: TheThe *histoghistogram* i is the simplest and m familiarfamiliar examplefamiliar example of afamilia = probability distribution function).

Figure 3. Normal and Lognormal Data from Figures 1 and 2 Displayed Using Histograms

Displays of **EDA**

WhatWhat can we saWhat can we say abouWhat can histograms in Figure 3?

Figure (a) appears to be bimodal.

Figure (b) is definitely right skewed.

Note: HistogramsHistograms can give different vis
impressionsimpressions timpressions thaimpressi
arbitraryarbitrary choicearbitrary choice of thearbitrar
ofof the intervals. The choice determines
whetherwhether we retain smoothness and simplicity
(a) or show more detail (b).

Displays of **EDA**

TheThe tradeoff between smoothness and closenessThe tradec datadata is dedata is determineddata is determined by t thumb thumb for determining the bin width (*h*)) where t) reference density is the Normal are:

h_1 = {range ($\boldsymbol{y}$) / $\log_2 n$ + 1}, Sturges formula,

h_2 = {3.5 $n^{1/3}$}, Scott (1979), and

h_3 = {2 IQR $n^{1/3}$}, Freedman & Diaconis}, Freedman &

wherewhere $\boldsymbol{y}$ is the sample vec is the sample vect is the and is the estimated standard deviation of $\boldsymbol{y}$.

Note: Take a look at the web page
http://www.stat.sc.edu/~west/javahtml/Histogram.html
for an applet on histograms and bin width.

HerbertHerbert Sturges (1926) was the fHerbert Sturges sysystesystematicsystematic guidelines for designing a histog oobobservedobserved that the binomial distribution, B(n, p coulcouldcould be used as a model of an optimallo constructed histogram.

ConstructConstruct aConstruct a frequency histogramCon withwith width 1 centered on the point i = *0, 1, 2, ...,* k *1*.

ChooseChoose the bin cChoose the bin count Choose Binomial coefficient

.

The total sample size is

.

By the Binomial expansion Sturges rule follows

$$k = 1 + \log_2 n.$$

InIn the caIn the case ofIn the case of Scott and F&D, the ru compromisescompromises betwcompromises betwecompro histogramhistogram using the Normal as the referen distribution.distribution. The bin width, *h,* is viewed a smoothing parameter.

TheThe *variance* can be reduced can be reduced by makin thethe bithe bins arthe bins are wide and approximate height.height. The *variance* can be eliminat can b choosingchoosing h = range($\boldsymbol{y}$) (all) (all observations are 1

TheThe *bias* can can be reduced can be reduced by makin bibinsbins bins are narrow. The bias can be eliminated choosingchoosing h = {min|y_i y_j|, wh|, where i j observation is its own bin).

Definition . . .

IIfIf the mIf the mean of all possible values of a statistic eqequalequal to a parameter, the statistic is called equ *unbiased estimator* of that parameter.

For example,

TheThe sampleThe sample mean, , is, is an *unbiased estim* populationpopulation mean, , bec, becaus, because t mmeansmeans of all possible samples of a givenmeans equal to the population mean.

Definition (cont.) . . .

Example of the Mean as an Unbiased Estimator

Population of 3 observations (2, 4, 6) where = 4

Sample No.	All possible *samples* (x_1, x_2)	Mean — = $(x_1 + x_2)/2$
1	2, 2	2
2	2, 4	3
3	2, 6	4
4	4, 2	3
5	4, 4	4
6	4, 6	5
7	6, 2	4
8	6, 4	5
9	6, 6	6
Sum —		36
Mean — /9		4

Definition . . .

LetLet y_1, y_2, y_3, ..., y_n be a be a rand be a random san
probabilityprobability distribution depends on an unkno
parameter,parameter,   . Let ,   = *f* (y_1, y_2, y_3,, ., ..., y_n) be
statisticstatistic (for examstatistic (for example, statistic (
squared error (MSE) of , as an estimator for is

$$\text{MSE}(\quad) = \text{Var}(\quad) + [\text{Bias}(\quad)]^2,$$

wwhewherewhere [Bias( )]^2 = [E[ ]   ]^2 and E repres
the *expected value*.

TheThe *bias* a and *variance* are controlled by choosing and intermediateintermediate value between {min|y_i y_j|, wher j,, range($\boldsymbol{y}$)} and allowing the bin w)} and allowing the b decrease as the sample size increases.

Definition . . .

AA density estimator isA density estimator is saidA dens *mean square error* if the

as n .

AnAn *optimal smoothing parameter*,, h^*, is defined to be that choice that minimizes the MSE.

Table 1. Comparison of the Number of Bins from the Three Normal Reference Rules*

Number of Bins	***Sturges*** $[\log_2 n + 1]$	***Scott*** $\left[\frac{\quad}{\quad}\right]$	***F&D*** $\left[\frac{\quad}{\quad}\right]$
50	5.6	6.3	8.5
100	7.6	8.0	10.8
500	10.0	13.6	18.3
1,000	11.0	17.2	23.2
5,000	13.3	29.1	39.6
10,000	14.3	37.0	49.9
100,000	17.6	79.8	107.6

** Scott, D. W. (1992) *MultivariateMultivariate Density Estimation. Multivariate De andand Visualization.* New York: John Wil New York: John Wiley New Yor datadata used to estimate the bin numbersdata used to estimate the bin numb 3).

Some observations about Table 1:

TheThe rules are comparable for The rules are compar than 100.

ForFor sample sizes greater thanFor sample sizes greater willwill prowill provwill provide an oversmoothed h waste much of the information in the data.

TheThe Freeman-Diaconis rule haThe Freeman-Diacon thanthan Scott than Scott sthan Scott s rule and theref smooth histogram.

Comparison of the Number of
Bins from the Three Normal Reference Rules
for the Data in Figures 1 and 2

Number of Bins	***Sturges***	***Scott***	***F&D***
Normal	7	4	5
Lognormal	7	6	32

Displays of **EDA**

A *cool* example that illustrates the power of EDA.

ForFor the LANL Environmental RestoratFor the LANL Envi extensive site characterization was performed.

SurfaceSurface soil samples are collected and compared to LANLLANL backgrouLANL background datLANL backgr samplessamples were collected from MDAsamples were collec elemental uranium.

MDA G Uranium Concentration (mg/kg)

Displays of EDA

WhatWhat can weWhat can we see in these data? What can w

ContamiContaminaContaminationsContaminations dif historical disposal operations.

GeologyGeology issues - diffeGeology issues - diffe differentdifferent background concentration ddifferent ba

ChemistryChemistry issues - different analytical metho data comparability issues.

SampleSample collection issues - different sampling methods,methods, differentmethods, different field team, issues.

MDA G Uranium Concentration (mg/kg)
by Analytic Technique

The problem was one ofThe problem was one of lack ofThe p inclusion of two very different analytical methods.

EDAEDA plots brought about a change in policy, KPA EDA nono longer beino longer being useno longer being used LANL ER Project.

Definitions . . .

AA *theoretical quantile-quantile plot (Q-Q plot) or probabilityprobability plot* is ob is obtained is quantilesquantiles of the observed data against th correspondingcorresponding quantiles ofcorrespor distribution (for example, the normal).

How do you construct a normal *Q-Q plot*?

Let $y_{1,}$, y_2, y_3, ..., y_n represent the raw data:

sort the data from smallestsort the data from smallest to t
$y_{(3)}$, ..., $y_{(n)}$,

calculatecalculate the empirical quancalculate the
observation, $Q_e(p_i)$, where

$$p_i = (i \quad 0.5)/n$$

((i.e.,(i.e., for a sample size of n = 20, the fifth sma = 20, th
observation,observation, $y_{(5)}$, is the {(*5* 0.5)/*20*}th quanti
$Q_e(0.225)$),

calculatecalculate the corresponding quantiles for the
standardstandard normal distribution (= = 0, = 1), if = 1), i
cumulativecumulative distribution function of the standa
normal, then

$$Q_t(p_i) = F\ ^{1}(p_i).$$

Let s try an example.

Table 2. Simple example for constructing a Standard Normal *Q-Q plot*

$y_{(i)}$	$p_{(i)}$	$Q_e(p_i)$	$Q_t(p_i)$*
7	0.05	7	1.64
8	0.15	8	1.04
11	0.25	11	0.38
13	0.35	13	0.13
14	0.45	14	0.13
17	0.55	17	0.38
18	0.65	18	0.38
19	0.80	19	0.84
19	0.80	19	0.84
20	0.95	20	1.64

* TheseThese valuThese values These values are found by looking them up ir tabletable or using a software packagetable or using a software package that quantiles.

IfIf the quantiles of the empirical distribIf the quantiles of andand theand the quantiles ofand the quantiles of theoretical on a straight line then the distributions are similar.

We can think of this another way.
Let,

$$F(y) \quad = \quad \left(\frac{\quad}{\quad} \right) \quad = \quad G(z)$$

where,

$z \quad = \quad \frac{\quad}{\quad}$ is the standardized variable and

G(z) is the CDF of the is the CDF of the
variable *Z*.

$z = G^{1}(F(y)) = \frac{\quad}{\quad} = \frac{}{} \quad \frac{}{}$

or in terms of *y* on *z*,

$y = \quad + \quad z.$

WhatWhat we are doing is trWhat we are doing is tran valuesvalues to standard normal variates. values to standard r *Q plot* the intercept is and the slope is .

Constructing the Standard Normal Q-Q plot from the data in Table 2.

TheThe Standard Normal Q-Q plot for the dataThe Standard N

Properties of the theoretical *Q-Q plot*:

IfIf the If the theIf the theoretical distribution
approximationapproximation to thapproximation to the
points on the plot will fall near the $y = x$ line.

IfIf the points follow a line that isIf the points follow a line
x li line, then the appropriate positive or negati line, the
constantconstant could be added to all dataconstant could
the configuration onto the $y = x$ line.

↓

Conclude: The empirical dist The empirical
compatiblecompatible with the theoreticalcompatible wi
theythey have diffethey have differentthey have diffe
means or medians.

Properties of the theoretical *Q-Q plot* (cont.):

IIfIf the points follow a line that is nearly straigIf the p
andand pass through the originand pass through the origi
thethe y = = x line, line, then it is possible to fin
appropriateappropriate positive constant by which
multiplymultiply all observations to multiply all observa
thethe configuration vertically and shift tthe co
configuration onto the $y = x$ line.

↓

Conclude: The empiric The empirica The
compatiblecompatible with the compatible with the the
theythey have different sprethey have different spreads
standard deviation or interquartile range.

Displays of **EDA**

Properties of the theoretical *Q-Q plot* (cont.):

TheThe straightness of theThe straightness of the theoretic judgejudge whether the empirical and rjudge wh distributiondistribution have the same distributional sha shiftsshifts and tshifts and tilts awashifts and tilts away differences in location and spread, respectively.

AA single theoretical *Q-Q plot* compares a set of datadata not just to one theoretical distribdata not just simultaneouslysimultaneously tosimultaneously to a who wiwithwith different locations (means) and spreawit (standard deviations).

HowHow do weHow do we interpret aHow do we interpret a t are deviations from the straight line pattern?

NotNot only does the *Q-QQ-Q plot* provide a warning provide matchmatch is poor, but it may alsomatch is poor, but it may al mismatch.

WhenWhen there are departures from linearity in a *Q-Q plot* theythey frequently match one of the following descriptions:

1. outliers at either end,
2. curvature at both ends,
3. convex or concave curvatures, and
4. horizontal segments, plateaus, or gaps.

Departures from linearity:

1. *outliers at either end,*
   AreAre the most extreme observations even larger thanthan couldthan could be reasonably expected for sam this size from the distribution in questions?

   TheThe theoretical *Q-Q plot* provides provides an informa effective answer.

Departures from linearity:

2. *curvature at both ends,*
   AnAn indication thAn indication theAn indication longerlonger or shorter tails than the theorelonger distribution.
      S-shaped S-shaped S-shaped first abo S-shap line,line, indicates heavline, indicates heavierline, indi distribution.
      S-shaped S-shaped first below then above the line,line, indicatesline, indicates lighter tails than the t distribution.

Departures from linearity:

3. *convex or concave curvatures,*
   AnAn indiAn indicatAn indication the theoretical symmetric and the empirical one is not.

   C-shaped C-shaped (concave) below the y=x l indicates positively skewed data.

   C-shaped C-shaped (convex) ab C-shaped (c indicates negatively skewed data.

Departures from linearity:

4. *horizontal segments, plateaus, or gaps*
   GraGranularityGranularity in the data which occurs at c
   valuesvalues may be due to rounding (horivalues
   segments).segments). Plateaus or gaps may be ansegment
   of more than one theoretical distribution.

Cautions for interpreting theoretical *Q-Q plots*:

1.1. TheThe natural variability ofThe natural variability of the didistributionaldistributional distributional modeldistribu from straightness.

Cautions for for interpreting theoretical for interpreting theor

2. EachEach *Q-Q plot* on only only compares the on
distributiondistribution of onedistribution of one distr
distribution;distribution; distribution; alldistribution; all d
datadata set, in particular the relationshipdata set, in
variable to others is ignored.

↓

Conclusion: Theoretical *Q-Q plQ-Q plots* are not a
panaceapanacea anpanacea and mupanacea and must
otherother displaysother displays and analyses to get a fu
the behavior of the data.

CDA *Introduction*

What is *confirmatory data analysis* (CDA)?

TheThe role of CDA is closer to that of traditional ststatisticalstatistical inference. It provides statementss significancesignificance and confidence, for example, inferen goodness-of-fitgoodness-of-fit tests and tests forgoodness-of- it sit s function is to provide the statistician with insight intointo ainto a set of data prior tointo a set of data prior to eva the investigation and drawing conclusions.

CDA methods are used to assess

thethe *reproducibility* of observed patterns or effec of obse and

goodness-of-fit using statements using statements of confid significance.

Goodness-of-Goodness-of-fitGoodness-of-fit tGoodness-of-
hypothesishypothesis that a given hypothesis that a given
stated probability law *F(x)*.

The null hypothesis can be a *simple hypothesis*

whenwhen *F(x)* is com is completely specified, for exa
normalnormal with meannormal with mean, =
deviation, = 10; or

thethe null hypothesis can bthe null hypothesis can be a
whenwhen *F(x)* is not completely specified, is not complete
F(x) is normal with unspecified and .

Goodness-of-fitGoodness-of-fit tests for nGoodness-of-fit te into five categories:

chi-square tests,

empirical distribution function (EDF) tests,

moment tests,

regression tests, and

miscellaneous tests.

CDA *Goodness-of-fit Tests*

ItIt is verIt is very It is very hard to compare goodness-of-f
establishestablish criteria asestablish criteria as to whatestabl
particularparticular situation. Thparticular situation. Thepart
thatthat the alternative hypothesis for GOFthat the alternativ
vaguevague,vague, vague, for example the empirical distrib
normal.

ComparComparisonComparisonsComparisons between GO
power as the criteria.

Definition . . .

TheThe *power* of a test, **1** , is the p, is the prob, is t
rejecting the null hypothesis when it is in the null hypothes

Make the DECISION:	The NULL HYPOTHESIS is:	
	True	False
Not to Reject the Null Hypothesis	*Correct Decision* (1)	*Incorrect Decision* Type II Error ( )
to Reject the Null Hypothesis	*Incorrect Decision* Type I Error ( )	*Correct Decision* Power (1)

CDA *Goodness-of-fit Tests*

CChi-sChi-square type GOF tests were developed by Karl Pearson.

TheThe mechanics consisThe mechanics consist hypothypothehypothesizedhypothesized distribution (with parameters)parameters) into a multinomial distribution wi cells,cells, cells, countingcells, counting cells, counting thcells, eacheach cell and contrasting these, using a chi-square or likelihoodlikelihood ratio test statistic, likelihood ratio number observations for each cell.

CDA *Goodness-of-fit Tests*

SomeSome prominent chi-square GSome prominent chi-squa
Fisher, Watson-Roy, and Rao-Robson.

Recommendations:
ItIt is recommended that the chi-square GOF test not be
usedused in testing departures from normality when the
datadata are *complete*. (Complete data are. (Complete data ar
valuevalue ofvalue of each observation is observed. Anvalue
incompleteincomplete or censored data would be an analy
measurementmeasurement thameasurement that wasmeasu
limit.limit. limit. Here we don tlimit. Here we don t have com
thatthat the value ithat the value is less that the value i
valuevalue.)value.) value.) Other procedures to be discusse
powerful (D Agostino, 1986).

EmpiricalEmpirical distribEmpirical distributionEmpirical distribu
thethe dithe discrthe discrepancy between the EDF an
distributiondistribution function, and are useddistribution fu
thethe sample tothe sample to the dithe sample to the distrib
bebe complbe complebe completely specified or may conta
which must be estimated from the sample.

The EDF is $F_n(y)$ defined by

ForFor any x, $F_n(y)$ records the records the proportion of observ
llessless thaless than or equal to x. $F_n(y)$ is used to estimate F
InIn fact it is a consistent estimator of $F(y)$, since, since as, since
| $F_n(y)(y)$ $F(y)$ | decreases to zero with probability | decrease

CDA *Goodness-of-fit Tests*

TheThe EDF is jusThe EDF is just anoThe EDF is distributiondistribution of a random variable.distribution o empiricalempirical *cumulative relative frequency* which is simplesimple esimple example of a *cumulative distribution f* (CDF).

Definitions . . .

Frequency is the number of observations is the numb particularparticular class. The *relativerelative frequency* is fr expressedexpressed as a proportionexpressed as a propo frequency.frequency. frequency. This simplestfrequency. didistributiondistribution function (PDF) is the re frequency histogram.

TheThe cumulative frequency is the number of observationobservation less than or equal toobservation TheThe *relative cumulative frequency* is cumulat is cun frequencyfrequency expressed as a proportion or percent the total frequency.

Examples of a simple PDF and CDF.

EEDFEDF tests are based on the largest vertical diffeEDF tes
betweenbetween $F_n(y)$ and $F(y)$. They are . They are divic
classes, *supremum* and *quadratic*.

Supremum:

TheThe most well-known EDThe most well-known EDI
waswas introducedwas introduced by Kolmogorov inwas
isis referred to as the Kolmogorov-Smirnovis referred to a
thethe KS Test. *D* is the largest of two is the larg
differences:

1. $F_n(y) > F(y)$, $D^+ = sup_y\{ F_n(y) \quad F(y) \}$, and
2. $F_n(y) < F(y)$, $D \quad = sup_y\{ F(y) \quad F_n(y) \}$.

Combined we have,

$D = sup_y \mid F_n(y) \quad F(y) \mid = \max\{ D^+, D \quad \}$.

Graphical Representation of the KS Test

Quadratic:

TheThe The sThe second class of EDF gives weights, (the squared differences $[\, F_n(y) \quad F(y)\,]^2$.

OneOne example is the *CramCramer-voCramer-von Mis* W^2. For For the Cramer-von Mises statistic, $(y)) = 1$.

AnotherAnother example is theAnother example is the *An* A^2.. For the A. For the Ander. For the Anderson-D $[(F(y))(1 \quad F(y))]^{\ 1}$.

Recommendations:

TheThe most powerful EDF test apThe most powerful Anderson-Darling,Anderson-Darling, A^2. Power studies are prproviprovideprovide information on comparisons to non-EI tests.

ForFor testing normality, it is recommFor testing normali Kolmogorov-SmirnovKolmogorov-Smirnov (K-S) teKolmog only only a historical curiosity. only a historical curiosit ppowerpower ipower in comparison to other pr (D Agostino, 1986).

Moment type GOF tests can be regarded type GOF tests c beenbeen initiated by Karl Pearson. He recognized that deviationsdeviations from normality could be characterized c thethe the stthe standard third and fourth moments distribution.

A test of the third standardized moment $\sqrt{\quad\quad}$,

$$H_o\text{:}\ \sqrt{\quad\quad}\quad 0.$$

S_U Approximation (D Agostino, 1970)

(1) Compute $\sqrt{\quad\quad}$ from the sample data.

(2) Compute

$$\sqrt{\quad}\left[\frac{\quad\quad\quad\quad}{\quad\quad\quad\quad}\right]$$

$$\frac{\quad\quad\quad\quad\quad\quad}{\quad\quad\quad\quad\quad\quad}$$

$$\sqrt{\quad\quad\quad}$$

S_U Approximation (cont.)

(3) Compute

Z is approximately a standard normal variate.

This transformation is applicableThis transformation is applic
greatgreatergreater than or equa√l to eight.   test is excel
for detecting nonnormality dues to skewness.

A test of the fourth standardized moment b_2,

$$H_o: \quad _2 \quad 3.$$

Anscombe and Glynn Approximation (1983)

(1) Compute b_2 from the sample data

(2) Compute the mean and variance of b_2

(4) Compute the third standardized moment of b_2

Anscombe and Glynn Approximation (cont.)

(5) Compute

$$\frac{}{\sqrt{}}\left[\frac{}{\sqrt{}}\qquad\sqrt{\qquad\frac{}{}}\right]$$

(6) Compute

$$\left[\left(\qquad\frac{}{}\right)\ \left(\frac{}{\sqrt{}}\right)\quad\right]\sqrt{}$$

Z is approximately a standard normal variate.

TheThe b_2 test is primarily used to detect nonno test is prim duedue to due to nonnodue to nonnormal kurtosis thickness.

AA number of rA number of reA number of researchers hav these tests to produce an *omnibus* test of normality.

One *omnibus* test of normality is the *R*-test.

TheThe *R*-test is the simplest omnibus test, it cotest is the si
performing the

$\sqrt{\qquad}$ test at the $_1$ level of significance, and
the b_2 test at the $_2$ level of significance.

TheThe overalThe overall leveThe overall level of signifi
employs the Bonferroni s inequality, $_1$ + $_2$.

TheThe term *R*-test wastest was giventest was given to this pr
cacancan be viewed as employing rectangular coordinacan b
for the rejection of normality.

AnotherAnother *omnibus* test of normality is the K^2 te test D Agostino and Pearson (1973).

D Agostino and Pearson suggested the test statistic,

$$\sqrt{\quad}\qquad\qquad ,$$

asas an as an omnias an omnibus test where a standardizedstandardized norstandardized normstandardiz square variable with 2 degrees of freedom.

Recommendations:
The K^2 test is more powerful than the *R*-test.

Regression and *correlation* t type type GOF tests make use of orderorder statistics,order statistics, $y_{(i)}$. A. A straight line is aa *Q-Q plot* and GOF tests are constr and GOF tests are co statistics associated with the line,

$$E(y_{(i)}) = \quad + \quad m_i, \qquad (1)$$

wherewhere is a location parameter, is a scale parameter,parameter, anparameter, and m_i represents t distribution.

ThereThere are three main apprThere are three main approa the data fit equation (1).

1. A test based on the correlation coefficient.

2. A test based on the sum of squared residuals { },}, where . In order to provideprovide a scale-free test provide a scale-free test divided by another quadratic form.

3. TheThe scalThe scale parameterThe scale parameter, squaredsquared valuesquared value compared with anoth

2.

TheThe Shapiro-Wilk GOF test is based on The Shapiro-
method of testing the fit of model (1).

TheThe sThe steps for conducting the Shapiro-Wilk GOF teT
are provided below.

1. Calculate

   , where

   $r = (n \quad 1)/2$ if n is odd and $r = n/2$ if n is even,
   andand a_i s are the optimal s are the optimal weights for
   leastleast squares estimator of ,, giv, given that t
   population is normally distributed.

2. Calculate

   $W = Y^2 / S^2$.

3. IfIf W is less than the value in the lower t is less than the
   tatabletable for the percentage points of the W-te-test fo
   normality,normality, for a particunormality, for a parti
   null.

TheThe exactThe exact distributionThe exact distribution of *W* dependsdepends on *n*. Since this distribution is not . Sin ShapiroShapiro and Wilk provided MonteShapiro and Wilk pointspoints for use with the test for points for use with the 50.

Recommendations (D Agostino, 1986):

GraphicalGraphical analyses should *always* ac accom ac formal test for normality.

TheThe Shapiro-Wilks *W* test and the D Agost test and PearsonPearson K^2 test appear to be the test appear to be th availaavailabavailable.available. The Shapiro-Wilks t probably overall most powerful.

TheThe K-S test should never be used. It hThe K-S tes power in comparison to other procedures.

WhenWhen tWhen tesWhen testing for normality with thethe chi-square test should nthe chi-square test should poor power in comparison to other procedures.

References

D Agostino,D Agostino, R. B., and Stephens, M. A. (1986). *Goodness-ofGoodness-of FGoodness-of Fit T* Marcel Dekker.

D Agostino,D Agostino, R. B. (1986). Tests for D Agostino, R. B. (1986). Tests for the ND A *Techniques*, D Agostino, R. B., and Stephens, M. A., Eds., pages 367 - 419.

D Agostino,D Agostino, R. B., Belanger, A., andD Agostino, R. B., Belanger, A., and D Agostino, Jr., R. PowerfulPowerful and Informative Tests of Normality.Powerful and Informative Tests of Norma

Freedman,Freedman, D. and Diaconis, P.Freedman, D. and Diaconis, P. (1981). On the histogramFreed *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**, 453-476.

Hand,Hand, D.Hand, D. J. (1999). Chapter 1. Hand, D. J. (1999). Chapter 1. Introduction. In *Intelligent* Hand D. J., Eds., pages 1-14.

Kruskal, W. H. (1960). Some remarks on wild observations. *Technometrics* **2**, 1-3.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66**, 605-610.

Scott,Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice, and Visualization.* New York: John Wiley and Sons.

Tietjen,Tietjen, G. L. (1986). Tietjen, G. L. (1986). The AnalysisTietjen, G. L. (1986). The Analysis and De D Agostino, R. B., and Stephens, M. A., Eds., pages 497 - 522.

Wainer, H. (1984). How to Display Data Badly. *The American Statistician* **38**, 137-146.